Queueing Systems
https://doi.org/10.1007/s11134-020-09649-9

Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue

Yan Chen[1] · **Ward Whitt**[1]

Received: 30 April 2019 / Revised: 13 February 2020

Abstract
It has long been conjectured that the tight upper bound for the mean steady-state waiting time in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions is attained asymptotically by two-point distributions. The two-point distribution for the interarrival time has one mass point at 0, but the service-time distribution involves a limit; there is one mass point at a high value, but that upper mass point must increase to infinity while the probability on that point must decrease to 0 appropriately. In this paper, we develop effective numerical and simulation algorithms to compute the value of this conjectured tight bound. The algorithms are aided by reductions of the special queues with extremal interarrival-time and extremal service-time distributions to $D/GI/1$ and $GI/D/1$ models. Combining these reductions yields an overall representation in terms of a $D/RS(D)/1$ discrete-time model involving a geometric random sum of deterministic random variables (the $RS(D)$), where the two deterministic random variables in the model may have different values, so that the extremal steady-state waiting time need not have a lattice distribution. Efficient computational methods are developed. The computational results show that the conjectured tight upper bound offers a significant improvement over established bounds.

Keywords Single-server queue · Bounds for the mean waiting time · Extremal queues · Stochastic simulation · Two-point distribution

Mathematics Subject Classification Primary: 60K25 · Secondary: 65C50, 90B22

✉ Ward Whitt
ww2040@columbia.edu

Yan Chen
yc3107@columbia.edu

[1] Department of IEOR, Columbia University, New York, NY, USA

Published online: 29 February 2020

1 Introduction

An open problem for the $GI/GI/1$ queue is to determine a tight upper bound (UB) for the mean steady-state waiting time when the interarrival-time and service-time distributions are partially characterized by their first two moments. The classic result in this direction is the Kingman [14] UB. A significant improvement was made by Daley [10]. Many further studies have been made since then, as can be seen from [11,29], but the issue is still unresolved.

Based on numerical experiments and partial theoretical results, there is considerable agreement on how the tight UB arises. In particular, the conjectured upper bound is attained asymptotically by two-point distributions as the upper mass point of the service-time distribution increases and the probability decreases, while one mass of the interarrival-time distribution is fixed at 0. In this paper, we shall refer to that as the tight UB, even though it has not yet been proved. We develop effective numerical and simulation algorithms to compute the value of this tight UB. We show that the tight UB offers a significant improvement over previous bounds.

The extremal model is also of interest because we have shown in [6,7] that it yields the tight lower bound (LB) for the asymptotic decay rate and the tight upper bound (UB) for higher moments of the steady-state waiting time, under regularity conditions. (The LB for the decay rate is appropriate because smaller decay rates are associated with larger waiting times.)

The algorithms are aided by reductions of the special queues with extremal interarrival-time and extremal service-time distributions to $D/GI/1$ and $GI/D/1$ models. Theorem 3.1 shows that these reductions overall combine to yield an overall representation in terms of a $D/RS(D)/1$ discrete-time model involving a geometric random sum of deterministic random variables (the $RS(D)$), where the two deterministic random variables in the model may have different values, so that the extremal steady-state waiting-time distribution need not be a lattice distribution.

This paper is organized as follows: In Sect. 2, we give important background on the $GI/GI/1$ model, the established bounds and the extremal distributions. In Sect. 3, we state the overall model reduction theorem and show the improvement provided by the tight UB. In Sects. 4 and 5 we establish the two reductions. In Sect. 6, we give explicit formulas for a class of special cases. In Sects. 7–9, we develop and study the new algorithms. In Sect. 10, we draw conclusions. Additional supporting material appears in [5].

2 Background

We start by carefully defining the model, reviewing established bounds and introducing the extremal distributions.

2.1 The model

The $GI/GI/1$ model is a single-server queue with unlimited waiting space and the first-come first-served service discipline. There is a sequence of independent and identically distributed (i.i.d.) service times $\{V_n : n \geq 0\}$, each distributed as V with cumulative distribution function (cdf) G, which is independent of a sequence of i.i.d. interarrival times $\{U_n : n \geq 0\}$ each distributed as U with cdf F. With the understanding that a 0^{th} customer arrives at time 0 to find an empty system, V_n is the service time of customer n, while U_n is the interarrival time between customers n and $n+1$.

Let U have mean $\mathbb{E}[U] \equiv \lambda^{-1} \equiv 1$ and squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2; let a service time V have mean $\mathbb{E}[V] \equiv \tau \equiv \rho$ and scv c_s^2, where $\rho \equiv \lambda\tau < 1$, so that the model is stable. (Let $\equiv$ denote equality by definition.)

Let W_n be the waiting time of customer n, i.e., the time from arrival until starting service, assuming that the system starts empty with $W_0 \equiv 0$. The sequence $\{W_n : n \geq 0\}$ is well known to satisfy the Lindley recursion

$$W_{n+1} = [W_n + V_n - U_n]^+, \quad n \geq 0, \tag{2.1}$$

where $x^+ \equiv \max\{x, 0\}$. Let W be the steady-state waiting time. It is also well known that $W_n \stackrel{\text{d}}{=} \max\{S_k : 0 \leq k \leq n\}$ and $W \stackrel{\text{d}}{=} \max\{S_k : k \geq 0\}$, where $\stackrel{\text{d}}{=}$ denotes equality in distribution, $S_0 \equiv 0$, $S_k \equiv X_0 + \cdots + X_{k-1}$ and $X_k \equiv V_k - U_k, k \geq 0$; for example, see Sects. X.1–X.2 of [3] or (13) in Sect. 8.5 of [8]. It is also known that, under the specified finite moment conditions, W_n and W are proper random variables with finite means, given by

$$\mathbb{E}[W] = \sum_{k=1}^{n} \frac{\mathbb{E}[S_k^+]}{k} < \infty \quad \text{and} \quad \mathbb{E}[W] = \sum_{k=1}^{\infty} \frac{\mathbb{E}[S_k^+]}{k} < \infty. \tag{2.2}$$

For numerical computation of $\mathbb{E}[W]$, formula (2.2) is unattractive, because it contains an infinite sum of terms and thus involves a k-fold convolution integral for all $k \geq 2$. Effective algorithms avoid that computational approach. One way to proceed is to apply numerical transform inversion with the Pollaczek contour integral representation, as in (5) of [2], i.e.,

$$\mathbb{E}[W] = \frac{1}{2\pi i} \int_C \log\{1 - \phi(z)\} \frac{\mathrm{d}z}{z}, \tag{2.3}$$

where $i \equiv \sqrt{-1}$, z is a complex variable,

$$\phi(z) \equiv \mathbb{E}[e^{z(V-U)}] \tag{2.4}$$

and C is a contour in the complex plane to the left of, and parallel to, the imaginary axis, and to the right of any singularities of $\log\{1 - \phi(z)\}$ in the left-half plane. As a regularity condition, we assume that the transform ϕ in (2.4) is analytic in the complex

plane for z in the strip $|z| < \delta$ for some $\delta > 0$. As in many probability applications, convolution is avoided by considering the transform in (2.4).

Unfortunately, our model with two-point distributions does not satisfy the regularity condition; for example, see Sect. 14 of [1]. As shown in [2], that difficulty can be avoided by asymptotic arguments. That was illustrated by calculating the cumulants and distribution of W in the $E_k/E_k/1$ for a wide range of k, even up to $k = 10^4$. In this paper, we will derive model reductions that will also enable us to avoid direct convolution in other ways.

2.2 The established upper bounds

Recall that our goal is to bound $\mathbb{E}[W]$ given a partial specification of the model, characterized by the parameter vector

$$(\mathbb{E}[U], \mathbb{E}[U^2], \mathbb{E}[V], \mathbb{E}[V^2]) = (1, c_a^2 + 1, \rho, \rho^2(c_s^2 + 1)). \tag{2.5}$$

Given that we have set $\mathbb{E}[U] = 1$, we have the parameter triple (ρ, c_a^2, c_s^2).

In this setting, the classical UB for $\mathbb{E}[W]$ is the Kingman [14] UB,

$$\mathbb{E}[W] \leq \frac{\rho^2([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)}. \tag{2.6}$$

An improvement is provided by the Daley [10] UB, which replaces the term c_a^2/ρ^2 by $(2-\rho)c_a^2/\rho$, i.e.,

$$\mathbb{E}[W] \leq \frac{\rho^2([(2-\rho)c_a^2/\rho] + c_s^2)}{2(1-\rho)}. \tag{2.7}$$

Both of these bounds are asymptotically correct in heavy traffic, i.e.,

$$\lim_{\rho \to 1} \frac{(1-\rho)}{\rho^2} \mathbb{E}[W] = (c_a^2 + c_s^2)/2, \tag{2.8}$$

which also supports the commonly used heavy-traffic approximation (HTA)

$$\mathbb{E}[W] \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)}. \tag{2.9}$$

The HTA reduces to the Pollaczek-Khintchine exact formula when the arrival process is a Poisson process, so that $c_a^2 = 1$. The heavy-traffic limit also shows that the scaled waiting-time distribution is asymptotically exponential and thus is asymptotically fully characterized by its mean.

2.3 The extremal distributions

In general, by "extremal model," we mean a model that yields the largest or smallest value of some performance measure given a partial model specification. We are considering the extreme values of $\mathbb{E}[W]$ in the $GI/GI/1$ queue with the parameters in (2.5). Thus, the extremal model means the pair of cdf's (F, G) of (U, V) yielding the extreme values of $\mathbb{E}[W]$.

2.3.1 Large service times versus large interarrival times

An intuitive explanation of the extremal distributions for $\mathbb{E}[W]$ in the $GI/GI/1$ queue is provided by considering the very different way that $\mathbb{E}[W]$ is affected by an exceptionally large interarrival time compared to an exceptionally large service time. A large interarrival time will tend to empty the queue seen by the next customer, but making it even larger has no further impact. On the other hand, a large service time increases waiting for many following customers, and making it even larger increases that impact. We elaborate in Sect. 2.4.

2.3.2 The two-point distributions

A special role is played by two-point distributions, which necessarily have finite support. Let $\mathcal{P}_{2,2}(m_1, c^2, M)$ be the set of all two-point distributions with mean m_1 and second moment $m_2 = m_1^2(c^2+1)$ with support in $[0, m_1 M]$. The set $\mathcal{P}_{2,2}(m_1, c^2, M)$ is a one-dimensional parametric family. Any element is determined by specifying one mass point. Let $F_b^{(2)}$ be the cdf that has probability mass $c^2/(c^2+(b-1)^2)$ on $m_1 b$, and mass $(b-1)^2/(c^2+(b-1)^2)$ on $m_1(1-c^2/(b-1))$ for $1+c^2 \le b \le M$. The cases $b = 1+c^2$ and $b = M$ constitute the two extremal distributions.

Since we are only interested in the extremal cdf's here, we will use different notation. We let $F_0 \equiv F^{(2)}_{1+c^2}$ because it is the unique element that has lower mass point 0, and we let $F_u \equiv F^{(2)}_M$ because it is the unique element that has upper mass point $m_1 M$. We use this definition for both the cdf's we consider: F of U and G of V, but recall that our parameter specification in (2.5) with $\mathbb{E}[U] = 1$ makes the support of F equal to $[0, M_a]$, while $\mathbb{E}[V] = \rho$ makes the support of G equal to $[0, \rho M_s]$. Therefore, with $M_a \ge 1 + c_a^2$ for F and $M_s \ge 1 + c_s^2$ for G, we have:

- $F_0 : c_a^2/(1+c_a^2)$ on 0, $1/(1+c_a^2)$ on $1+c_a^2$;
- $F_u : (M_a-1)^2/(c_a^2+(M_a-1)^2)$ on $1-c_a^2/(M_a-1)$, $c_a^2/(c_a^2+(M_a-1)^2)$ on M_a;
- $G_0 : c_s^2/(1+c_s^2)$ on 0, $1/(1+c_s^2)$ on $\rho(1+c_s^2)$;
- $G_u : (M_s-1)^2/(c_s^2+(M_s-1)^2)$ on $\rho(1-c_s^2/(M_s-1))$, $c_s^2/(c_s^2+(M_s-1)^2)$ on ρM_s.

2.3.3 Special notation

Given that we are interested in the pair of cdf's (F, G) of (U, V) yielding the extreme values of $\mathbb{E}[W]$, it will be convenient to introduce some special notation, which

departs from the classic Kendall queueing notation, without changing the independence assumptions. In particular, we will often denote a $GI/GI/1$ model with cdf pair (F, G) as the $F/G/1$ model or as the $U/V/1$ model (with the understanding that U and V are then shorthand for their cdf's). We might even mix the notation, for example, referring to the $D(m)/G/1$ model, where $D(m)$ represents a deterministic random variable with mean m. (The G without the GI refers to the service-time cdf G.) We will also refer to the steady-state waiting time as a function of the pair (F, G) as $\mathbb{E}[W(F, G)]$ or $\mathbb{E}[W(U, V)]$.

We also introduce some notation to account for the limit we consider, which involves $M_s \to \infty$ in the setting above. In particular, with a slight abuse of notation, for any cdf F of U, we let G_{u^*} denote the limit

$$\mathbb{E}[W(F, G_{u^*})] \equiv \lim_{M_s \to \infty} \mathbb{E}[W(F, G_u)]. \tag{2.10}$$

This definition is formalized and justified in Theorem 5.1.

2.4 Asymptotics for the extremal distributions

With this notation, we can elaborate on the impact of very large interarrival times and service times discussed in Sect. 2.3.1. For that purpose, let $\Rightarrow$ denote convergence in distribution; for example, see [4]. Let the cdf's F and G have parameters in (2.5).

In this setting, it is easy to see that these two-point distributions have deterministic limits as the support bounds increase; i.e.,

$$F_u \Rightarrow D(1) \quad \text{as} \quad M_a \to \infty \quad \text{and} \quad G_u \Rightarrow D(\rho) \quad \text{as} \quad M_s \to \infty, \tag{2.11}$$

even though the scv of the deterministic limit is 0 instead of the original value.

2.4.1 The tight lower bound

The tight lower bound (LB) for $\mathbb{E}[W]$ in $GI/GI/1$ subject to (2.5) has long been known; see [23], Sect. 5.4 of [22], Sect. V of [19,25] and Theorem 3.1 of [11]. The LB is not attained at two-point distributions. The LB is attained asymptotically by $F = D(1)$ as $M_a \to \infty$ and $G = A_3$, where A_3 denotes any three-point service-time distribution (with the given mean ρ) that concentrates all mass on nonnegative-integer multiples of the deterministic interarrival time. The tight LB has explicit formula

$$\mathbb{E}[W(D(1), A_3)] = \frac{\rho((1 + c_s^2)\rho - 1)^+}{2(1 - \rho)}. \tag{2.12}$$

We will show the LB in numerical comparisons, but not focus on it.

Consistent with (2.12), we have an associated asymptotic result. In particular, by Theorem X.6.1 of [3], for any cdf G with the parameters in (2.5),

$$W(F_u, G) \Rightarrow W(D(1), G) \quad \text{and} \quad \mathbb{E}[W(F_u, G)] \to \mathbb{E}[W(D(1), G)] \quad \text{as} \quad M_a \to \infty. \tag{2.13}$$

That partly explains the LB in (2.12).

2.4.2 A contrasting story for the upper bound

In contrast, there is a very different story for the UB. To explain, let V_{M_s} be a random variable with distribution G_u as a function of M_s. By the definitions, $\mathbb{E}[V_{M_s}] = \rho$ and $\mathbb{E}[V_{M_s}^2] = \rho^2(1 + c_s^2)$ for all M_s. By (2.11), $V_{M_s} \Rightarrow D(\rho)$ as $M_s \to \infty$, but $\mathbb{E}[V_{M_s}^{2+p}] \to \infty$ for all $p > 0$ and the family $\{V_{M_s}^2 : M_s \geq c_s^2 + 1\}$ is not uniformly integrable; see Sect. X.6 of [3] and pages 30-32 of [4]. Nevertheless, by Theorem 5.1, for any cdf F, $\mathbb{E}[W(F, G_u)]$ converges as $M_s \to \infty$, but the limit is typically strictly larger than $\mathbb{E}[W(F, D(\rho))]$. Definition (2.10) is intended to capture the true limit as $M_s \to \infty$. That limit is given in Theorem 5.1 in Sect. 5. See Corollary 5.3 to Theorem 5.1 for more details.

3 Summary of the results

We now summarize our main results. We first state our main reduction theorem. Then, we develop an explicit UB formula for $\mathbb{E}[W(F_0, G_{u^*})]$ in (2.10). Then, we show the advantage of the tight UB over previous UB's.

3.1 The overall reduction theorem

Our main purpose in this paper is to develop and evaluate algorithms to efficiently compute $\mathbb{E}[W(F, G_{u^*})]$ in (2.10). That is challenging because the large service time is a rare event. For example, standard Monte Carlo simulation with the Lindley recursion and the inverse method is not so effective for estimating $\mathbb{E}[W]$ accurately. We show that effective algorithms can be developed if we transform the problem, which we do through two model reductions.

In Sect. 4, we introduce our first model reduction. Drawing on [13] or [24], we show that, for any service-time cdf G, the mean waiting time in the $F_0/G/1$ model can be expressed in terms of the mean waiting time in an associated $D(m)/G/1$ model with a new service-time distribution involving a geometric random sum. Then, drawing on [11], in Sect. 5, we introduce a second model reduction. We show that, for any interarrival-time cdf F, $\mathbb{E}[W(F, G_{u^*})]$ in (2.10) can be expressed in terms of the mean waiting time in an associated model.

For the statement of our main decomposition result (to be proved in the next two sections), (i) let $D(m)$ be a deterministic random variable with a unit mass on m and (ii) let $RS(V, p)$ be a geometric random sum of i.i.d. random variables distributed as V, i.e.,

$$RS(V,\, p) \stackrel{\mathrm{d}}{=} \sum_{k=1}^{N(p)} V_k, \tag{3.1}$$

where $N(p)$ is a geometric random variable on the positive integers, having mean $\mathbb{E}[N(p)] = 1/p$, i.e., $\mathbb{P}(N(p) = k) \equiv p(1-p)^{k-1}$, $k \geq 1$.

Theorem 3.1 (overall decomposition of the upper bound) *For the $GI/GI/1$ model with extremal interarrival-time cdf F_0 and service-time cdf G_u,*

$$\begin{aligned}\mathbb{E}[W(F_0, G_{u^*})] &\equiv \lim_{M_s \to \infty} \mathbb{E}[W(F_0, G_u)] \\ &= \mathbb{E}[W(D(1/p),\, RS(D(\rho),\, p))] + \rho c_a^2 + \frac{\rho^2 c_s^2}{2(1-\rho)} \end{aligned} \tag{3.2}$$

for $p \equiv 1/(c_a^2 + 1)$.

Remark 3.1 *(functional form)* Notice that the first term in (3.2) is independent of c_s^2, while the second term is independent of c_a^2, but the functional dependence on the pair (ρ, c_a^2) is somewhat complicated because $p = 1/(1 + c_a^2)$. The interarrival time and service time in the $D(1/p)/RS(D(\rho),\, p))/1$ model have means $1 + c_a^2$ and $\rho(1 + c_a^2)$, but the distribution is more complicated. The variance and scv of the service time are given in Lemma 4.1.

Remark 3.2 *(nonlattice distribution)* Theorem 3.1 may explain the long-standing difficulty establishing the tight upper bound, because the $D(1/p)/RS(D(\rho),\, p)/1$ model is a discrete model, where the service times are necessarily multiples of ρ, while the interarrival times are multiples of $1/p = c_a^2 + 1$. Hence, the distribution of W is discrete nonlattice for most parameter pairs (ρ, c_a^2). □

In Sect. 6, we exhibit the steady-state distribution of $\mathbb{E}[W(F_0, G_{u^*})]$ for a class of special cases, but it involves solving an equation for a key parameter. With (3.2), we can apply basic convex stochastic order relations to obtain a convenient upper bound formula for $\mathbb{E}[W(D(1/p),\, RS(D(\rho),\, p))]$. For that purpose, let $\leq_{cx}$ and $\leq_{icx}$ denote convex stochastic order and increasing stochastic order, respectively, as in Sect. 9.5 of [20] or Sect. 1.5 of [18]. Also let $M(m)$ be an exponential random variable with mean m. Since $D(m) \leq_{cx} M(m)$, we deduce the following theorem. To obtain an explicitly computable formula using mathematical software, we exploit the Lambert W function, here denoted by w, characterized by $z = w(z)e^{w(z)}$ for all complex variables z; see [9].

Theorem 3.2 *In the setting of Theorem 3.1,*

$$W(D(1/p),\, RS(D(\rho),\, p)) \leq_{icx} W(D(1/p),\, M(\rho/p)), \tag{3.3}$$

so that well-known results for the $D/M/1$ queue yield

$$\mathbb{E}[W(F_0, G_{u^*})] \leq \frac{[2(1-\rho)\rho/(1-\delta)]c_a^2 + \rho^2 c_s^2}{2(1-\rho)}, \tag{3.4}$$

where $\delta \in (0, 1)$ *solves the equation*

$$\delta = \exp(-(1-\delta))/\rho)), \quad \textit{so that} \quad \delta = -\rho w(-\rho^{-1}e^{-1/\rho}), \tag{3.5}$$

where w *is the Lambert W function.*

Proof Since $D(m) \leq_{cx} M(m)$ for all m, we also have $RS(D(\rho), p) \leq_{cx} RS(M(\rho), p)$ by Theorem 9.6.7 of [20], but

$$RS(M(\rho), p) \stackrel{\mathrm{d}}{=} M(\rho/p), \tag{3.6}$$

as can be verified by computing the Laplace transforms. Then, for the transient waiting times, we have the ordering

$$W_n(D(1/p), RS(D(\rho), p)) \leq_{icx} W_n(D(1/p), M(\rho/p)) \quad \text{for all} \quad n \geq 1, \tag{3.7}$$

by Theorem 9.6.2 of [20]. Given our second moment conditions in (2.5), we also have the conclusion for the steady-state variables in (3.3) by Theorem 1.5.9 of [18].

Finally, we derive the explicit representation of the root δ using the Lambert W function w in the last line. Start with the equation $e^{bx} = cx$ for positive real numbers b and c, where x is a real variable. Then, let $y = -bx$, which implies that $-b/c = ye^y$. Now apply the characterization $z = w(z)e^{w(z)}$ to obtain $y = w(-b/c)$ and $x = -w(-b/c)/b$. Then, substitute for b and c. □

Remark 3.3 *(not tight)* Table 1 shows that the UB for $\mathbb{E}[W(F_0, G_{u^*})]$ in (3.4) is very accurate, but it is a not tight UB. We conjecture that formula (3.4) is a legitimate overall UB. Given Theorem 3.2, a sufficient condition is for $\mathbb{E}[W(F_0, G_{u^*})]$ to be the tight UB.

3.2 The algorithms

In Sect. 7, we use the representation in (3.2) to produce the first effective numerical algorithm involving the negative binomial distribution. For further progress, following [16,17,29], in Sect. 8, we review the representation of the mean waiting time $\mathbb{E}[W]$ in terms of the parameter vector $(1, c_a^2, \rho, c_s^2)$ and the idle-time distribution. When combined with the idle-time representation, this yields other convenient ways to calculate or estimate $\mathbb{E}[W]$ via numerical algorithms and simulations. We then study three simulation algorithms in Sect. 9.

3.3 The advantage of the tight UB

To show that the tight UB $\mathbb{E}[W(F_0, G_{u^*})]$ studied in this paper and its approximation formula in (3.4) provide significant improvements, we compare the numerical estimates of the tight UB for the $GI/GI/1$ model with given first two moments associated with $c_a^2 = c_s^2 = 4.0$ to other bounds and approximations in Table 1. Comparisons for

the cases $(c_a^2, c_s^2) = (0.5, 0.5)$, $(4.0, 0.5)$ and $(0.5, 4.0)$ appear in Sect. 2 of the online supplement, [5]. The heavy-traffic approximation (HTA) is (2.9). The MRE is the maximum relative error between the new UB in (3.4) and the conjectured tight UB $\mathbb{E}[W(F_0, G_u^*)]$ in (3.2).

Remark 3.4 *(need for additional information)* While Table 1 shows that the conjectured tight UB in (2.10) and (3.2) and its UB in (3.4) provide significant improvements over previous UB's, Table 1 also shows that there remains a wide range between the LB and the UB. That indicates that more reliable prediction depends on additional information. One way to address that is described in [6].

4 Reduction for the interarrival time

In this section, we show that for any service-time cdf G, the mean waiting time in the $F_0/G/1$ queue can be expressed in terms of the mean waiting time in an associated $D(m)/RS/1$ queue with a new service-time distribution involving a random sum. The key observation is that the $F_0/G/1$ queue corresponds to the $D/GI/1$ queue with batch arrivals; then, the new service-time cdf is the sum of the service times in the batch. However, we need to make other adjustments as well.

Let F_0 be the two-point upper bound extremal distribution with mean 1 and mass $p \equiv 1/(c_a^2 + 1)$ on $c_a^2 + 1$ and mass $1 - p$ on 0. Let $RS(V, p)$ be the random variable defined in (3.1). For the interarrival times, we will consider $D(x)$ for $x = 1/p = (c_a^2 + 1)$.

Theorem 4.1 *For any service-time V with cdf G having mean ρ and scv c_s^2, the steady-state waiting time is distributed as*

$$W(F_0, G) \stackrel{\mathrm{d}}{=} W(D(1/p), RS(V, p)) + \sum_{k=1}^{N(p)-1} V_k \tag{4.1}$$

for $N(p)$ and $RS(V, p)$ in (3.1), where the two terms in (4.1) are independent. Hence, the mean is

$$\begin{aligned} \mathbb{E}[W(F_0, G)] &= \mathbb{E}[W(D(1/p), RS(V, p)] + (\mathbb{E}[N(p)] - 1)\mathbb{E}[V] \\ &= \mathbb{E}[W(D(1/p), RS(V, p))] + \rho(1 - p)/p \\ &= \mathbb{E}[W(D(1/p), RS(V, p))] + \rho c_a^2. \end{aligned} \tag{4.2}$$

Proof The F_0 interarrival time means that a random number of arrivals, distributed as $N(p)$, arrive at deterministic intervals with deterministic value $1/p = c_a^2 + 1$. So the model has batch arrivals. The result in (4.2) follows from [13] or Theorem 1 of [24], which states that the delay of an arbitrary customer in the batch is distributed the same as the delay of the last customer in the batch when the batch-size distribution is geometric. Because $\mathbb{E}[W(D(1/p), RS(V, p))]$ is the expected delay of the first customer in a batch, we need to add the second term in (4.2) to get the delay of the last customer in the batch; for example, see Sect. III of [24]. □

Table 1 A comparison of the bounds and approximations for the steady-state mean $\mathbb{E}[W]$ as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$

ρ	Tight LB (2.12)	HTA (2.9)	Tight UB (3.2)	new UB (3.4)	δ (3.5)	MRE	Daley (2.7)	Kingman (2.6)
0.10	0.000	0.044	0.422	0.422	0.000	0.00%	0.444	2.244
0.20	0.000	0.200	0.904	0.906	0.007	0.19%	1.000	2.600
0.30	0.107	0.514	1.499	1.508	0.041	0.60%	1.714	3.114
0.40	0.333	1.067	2.304	2.326	0.107	0.94%	2.667	3.867
0.50	0.750	2.000	3.470	3.510	0.203	1.15%	4.000	5.000
0.60	1.500	3.600	5.295	5.352	0.324	1.07%	6.000	6.800
0.70	2.917	6.533	8.441	8.520	0.467	0.93%	9.333	9.933
0.80	6.000	12.800	14.917	15.017	0.629	0.67%	16.000	16.400
0.90	15.750	32.400	34.721	34.843	0.807	0.35%	36.000	36.200
0.95	35.625	72.200	74.621	74.755	0.902	0.18%	76.000	76.100
0.98	95.550	192.080	194.557	194.702	0.960	0.07%	196.000	196.040
0.99	195.525	392.040	394.533	394.684	0.980	0.04%	396.000	396.020

To work with the $D(1/p)/RS(V, p)/1$ model, we need the mean and variance of the random sum $RS(V, p)$ in (3.1).

Lemma 4.1 (random sum moments) *Given that V has mean ρ and scv c_s^2, the mean and variance of the random sum $RS(V, p)$ in (3.1) are*

$$\mathbb{E}[RS(V, p)] = \mathbb{E}[N(p)]\mathbb{E}[V] = \frac{\rho}{p} = \rho(c_a^2 + 1) \tag{4.3}$$

and

$$Var(RS(V, p)) = \rho^2 c_s^2(c_a^2 + 1) + \rho^2 c_a^2(1 + c_a^2). \tag{4.4}$$

Hence, the scv of $RS(V, p)$ is

$$\bar{c}_s^2 \equiv \frac{Var(RS(V, p))}{\mathbb{E}[RS(V, p)]^2} = \frac{\rho^2 c_s^2(c_a^2 + 1) + \rho^2 c_a^2(1 + c_a^2)}{\rho^2(1 + c_a^2)^2} = \frac{c_a^2 + c_s^2}{1 + c_a^2}. \tag{4.5}$$

Proof We apply the standard formulas for random sums from p. 113 of [21]. For the variance,

$$\begin{aligned} Var(RS(V, p)) &= Var(V)\mathbb{E}[N] + (\mathbb{E}[V])^2 Var(N) = \frac{\rho^2 c_s^2}{p} + \frac{\rho^2(1 - p)}{p^2} \\ &= \rho^2 c_s^2(1 + c_a^2) + \rho^2 c_a^2(1 + c_a^2), \end{aligned} \tag{4.6}$$

as claimed. □

Let $\bar{c}_a^2 = 0$ be the scv of $D(1/p)$ and recall that $p = 1/(1 + c_a^2)$.

Theorem 4.2 *For the $D(1/p)/RS(V, p)/1$ model, the Kingman [14] upper bound in (2.6) for the mean steady-state waiting time is*

$$\begin{aligned} \mathbb{E}[W] &\le \frac{\rho\mathbb{E}[RS(V, p)]((\bar{c}_a^2/\rho^2) + \bar{c}_s^2)}{2(1 - \rho)} \\ &= \frac{\rho^2(1 + c_a^2)\bar{c}_s^2}{2(1 - \rho)} = \frac{\rho^2(c_a^2 + c_s^2)}{2(1 - \rho)}. \end{aligned} \tag{4.7}$$

Hence, the associated upper bound for the $F_0/G/1$ model is

$$\mathbb{E}[W(F_0, G)] \le \frac{\rho^2(c_a^2 + c_s^2)}{2(1 - \rho)} + \rho c_a^2 = \frac{\rho^2(Ac_a^2 + c_s^2)}{2(1 - \rho)}, \tag{4.8}$$

where

$$A \equiv A(\rho, c_a^2) \equiv 1 + \frac{2(1 - \rho)}{\rho} = \frac{2}{\rho} - 1, \tag{4.9}$$

which makes (4.8) coincide with [10] bound in (2.7).

Proof We exploit Theorem 4.1, which provides the representation (4.2). Then, observe, with the aid of Lemma 4.1, that the [14] bound for $\mathbb{E}[W(D(1/p), RS(V, p))]$ is given by the first term on the first line of (4.8). □

5 Reduction for the service time

Daley proposed another decomposition that can be used to avoid the rare event of the large service time M_s. The Daley decomposition allows us to reduce the model $F/G_u/1$ to $F/D/1$ for arbitrary F as $M_s \to \infty$. The Daley decomposition is stated in (10.2) of the review paper [11] without proof, referring to an unpublished manuscript. As before, let $D(m)$ denote a deterministic cdf with mass 1 on m.

Theorem 5.1 (the Daley decomposition in (10.2) of [11]) *Consider the $F/G_u/1$ model with arbitrary interarrival-time cdf F and two-point service-time cdf G_u. Then,*

$$\lim_{M_s\to\infty} \mathbb{E}[W(F, G_u)] = \mathbb{E}[W(F, D(\rho))] + \lim_{M_s\to\infty} \mathbb{E}[W(D(1), G_u)]$$
$$= \mathbb{E}[W(F, D(\rho))] + \frac{\rho^2 c_s^2}{2(1-\rho)}. \tag{5.1}$$

Proof Given that this result is already known, we only outline our proof. We do a regenerative analysis to compute the mean waiting time, looking at successive busy cycles starting empty. We exploit the classic result that the steady-state mean waiting time is the expected sum of the waiting times over one cycle divided by the expected length of one cycle; for example, see Sects. 3.6 and 3.7 of [20].

As M_s increases, the two-point cdf G_u necessarily places probability of order $O(1/M_s^2)$ on M_s and the rest of the mass on a point just less than the mean service time, ρ. For very large M_s, there will be only rarely, with probability of order $O(1/M_s^2)$, be a large service time of order $O(M_s)$. In the limit, most customers never encounter this large service time, so that we get a contribution to the overall mean $\mathbb{E}[W]$ corresponding to $\mathbb{E}[W(F, D(\rho))]$ in the first term on the right in (5.1).

On the other hand, the total impact of the very large waiting time of order M_s is roughly the area of the triangle with height $O(M_s)$ and width $O(M_s)$, which itself is $O(M_s^2)$. When combined with the $O(1/M_s^2)$ probability, this produces an additional $O(1)$ impact on the steady-state mean, which is given by the second term on the right in (5.1). Moreover, because we can use a law-of-large-numbers argument to treat this large service time, the asymptotic impact of that large service time is independent of the interarrival-time cdf beyond its mean, so we can substitute $D(1)$ for the original interarrival-time cdf F with mean 1 in the second term.

To elaborate on the value of the last term, the mean cycle length is asymptotically 1. From the form of G_u in Sect. 2.3.2, we see that (i) the probability of the rare event is asymptotically c_s^2/M_s^2 and (ii) when the rare event occurs, the large service time of size ρM_s will arrive. Assuming no other large service times arrive thereafter, service times of size approximately ρ arrive every unit time, so that the queue will empty approximately at time $\rho M_s/(1-\rho)$. Hence, the sum of all the waiting times from the

arrival of the large service time until the queue first empties and the cycle ends is about $\rho^2 M_s^2/2(1-\rho)$. Putting these together, we see that the expected sum of all waiting times in the cycle is asymptotically $(c_s^2/M_s^2) \times \rho^2 M_s^2/2(1-\rho) = \rho^2 c_s^2/2(1-\rho)$, as stated. This reasoning can be made precise using the reasoning in [27]. □

Corollary 5.1 (decomposition of the upper bound) *For the $F/G_u/1$ model with G_{u*} in (2.10),*

$$\mathbb{E}[W(F_0, G_{u^*})] = \mathbb{E}[W(F_0, D(\rho))] + \frac{\rho^2 c_s^2}{2(1-\rho)}.$$

We can combine Theorem 4.1 and Corollary 5.1 to obtain Theorem 3.1. Corollary 5.1 implies that calculating the UB of $\mathbb{E}[W]$ is equivalent to calculating $F_0/D/1$, which has deterministic service time. Clearly, this makes the UB much easier to estimate by classical simulation methods.

Corollary 5.2 (tightness of Kingman's bound) *For the $D/G_u/1$ model with G_{u*} in (2.10),*

$$\mathbb{E}[W(D(1), G_{u^*})] = \mathbb{E}[W(D(1), D(\rho))] + \frac{\rho^2 c_s^2}{2(1-\rho)} = \frac{\rho^2 c_s^2}{2(1-\rho)},$$

so that Kingman's bound in (2.6) is asymptotically attained by the $D/G_u/1$ model as $M_s \to \infty$.

Finally, by the proof of Theorem 5.1, we also obtain the following negative result.

Corollary 5.3 (higher moments) *For the $F/G_u/1$ model and any $p > 0$,*

$$\lim_{M_s \to \infty} \mathbb{E}[W(F, G_u)^{1+p}] \to \infty \quad as \quad M_s \to \infty.$$

6 Explicit solution in special cases

From Corollary 5.1 in Sect. 5, we conclude that we can express the UB $\mathbb{E}[W(F_0, G_{u^*})]$ in terms of the mean $\mathbb{E}[W(F_0, D(\rho))]$. We now derive an explicit expression for this mean and the full steady-state waiting-time distribution in this $F_0/D(\rho)/1$ model for some special cases.

Theorem 6.1 *If $1 + c_a^2 = k\rho$ for a positive integer k, then*

$$\mathbb{E}[W(F_0, D(\rho))] = \frac{r\rho}{1-r}, \tag{6.1}$$

where $r \in (0, 1)$ is the root of $\sum_{l=1}^{k-1} x^l - c_a^2 = 0$. Moreover,

$$\mathbb{P}(W(F_0, D(\rho)) = l\rho) = (1-r)r^l, \quad l = 0, 1, 2, \ldots \tag{6.2}$$

Proof In the $F_0/D(\rho)/1$ model, the interarrival times and service times are lattice distributions, with the support of F_0 being $\{0, 1 + c_a^2\}$ and the support of $D(\rho)$ being $\{\rho\}$. Thus, under the condition $(1 + c_a^2) = k\rho$ for some integer k, $W(F_0, D(\rho))$ has support $\{l\rho : l \geq 0\}$. Let $p_l \equiv \mathbb{P}(W(F_0, D(\rho)) = l\rho), l \geq 0$. We see that p_l satisfies the recursion

$$p_{l+1} = p_l \frac{c_a^2}{(1 + c_a^2)} + p_{l+k} \frac{1}{(1 + c_a^2)}, \; p_0 = \left(\sum_{l=0}^{k-1} p_l \right) \frac{1}{(1 + c_a^2)}. \tag{6.3}$$

Given that $\{W_n : n \geq 0\}$ is an ergodic irreducible discrete-time Markov chain with a unique steady-state distribution, it suffices to find a solution to the recursion in (6.3). Thus, suppose that the p_l is a geometric pmf as in (6.2) with unknown r and p_0. We then see that there exist unique p_0 and r satisfying recursion (6.3). In particular, since $p_{l+1} = rp_l$ where $r \in (0, 1)$, we use the second equation of (6.3) to conclude that there exists a unique $r \in (0, 1)$ which is the unique root of the equation $\sum_{l=1}^{k-1} x^l - c_a^2 = 0$ with $p_0 = 1 - r$. □

7 The negative binomial numerical algorithm

In this section, we apply Theorem 3.1 to obtain an efficient algorithm for computing the UB $\mathbb{E}[W(F_0, G_{u^*})]$. Theorem 3.1 implies that it suffices to compute $\mathbb{E}[W]$ in the $D(1/p)/RS(D(\rho), p)/1$ model. The representation of the service time as a geometric random sum allows us to express $\mathbb{E}[W]$ directly in terms of the negative binomial (NB) distribution, without having to perform any convolutions.

Let $NB \equiv NB(n, p)$ be a conventional negative binomial random variable with parameter pair (n, p) for nonnegative integer n and $0 < p < 1$, which has probability mass function (pmf)

$$p_k(n, p) \equiv \mathbb{P}(NB(n, p) = k) \equiv \left(\frac{(n + k - 1)!}{k!(n - 1)!} \right) (1 - p)^n p^k, \quad n \geq 0, \tag{7.1}$$

with mean and variance

$$\mathbb{E}[NB(n, p)] = \frac{np}{1 - p} \quad \text{and} \quad Var(NB(n, p)) = \frac{np}{(1 - p)^2}. \tag{7.2}$$

Lemma 7.1 (NB representation of the mean) *For the* $D(1/p)/RS(D(\rho), p)/1$ *model,*

$$S_n \stackrel{\mathrm{d}}{=} \rho(NB(n, 1 - p) + n) - (n/p), \tag{7.3}$$

for S_n *in* (2.2)*, so that*

$$\mathbb{E}[W(D(1/p), RS(D(\rho), p))] = \sum_{n=1}^{\infty} n^{-1} \mathbb{E}[(NB(n, 1 - p) + n - (n/p))^+]. \tag{7.4}$$

To compute, $\mathbb{E}[W]$, we compute the transient mean $\mathbb{E}[W_N]$ in (2.2) for suitably large N, which means truncating the sum in (7.4). As often with the NB pmf, because of the factorials, it is convenient to use a recursive algorithm for computation. In the first version, we initialize the recursion at $k = 0$, letting $\mathbb{P}(NB(n, 1-p) = 0) = p^n$. Then, we can apply the recursion

$$\mathbb{P}(NB(n, 1-p) = k) = \frac{\mathbb{P}(NB(n, 1-p) = k-1)(n+k-1)(1-p)}{k}, \quad (7.5)$$

where $p = 1/(1 + c_a^2)$. However, for the parameter $p = 1/(c_a^2 + 1)$ already defined by F_0, we end up with negative binomial parameter $1 - p$. A recursive algorithm is given in Algorithm 1, with explanation afterward.

Algorithm 1 basic negative binomial recursion (k in outer loop)

Initially set $\mathbb{E}[W] \leftarrow \rho c_a^2 + \frac{\rho^2 c_s^2}{2(1-\rho)}$ and $p = (1 + c_a^2)^{-1}$.
for $k \in [1, K]$ **do**
  $S(k) \leftarrow 0, nbpdf \leftarrow p(1-p)^k$
  for $n \in [1, N]$ **do**
    $S(k) \leftarrow S(k) + nbpdf \max((n+k)\rho - n/p, 0)/n$
    $nbpdf \leftarrow nbpdf(\frac{n+k}{n})p$
  $\mathbb{E}[W] \leftarrow \mathbb{E}[W] + S(k)$
Output $\mathbb{E}[W]$

To explain Algorithm 1, recall that we are applying Theorem 3.1 to obtain an efficient algorithm for computing the UB $\mathbb{E}[W(F_0, G_{u^*})]$. Thus, we initialize by the constant term that depends only on the parameter vector $(1, c_a^2, \rho, c_s^2)$. We add that to $\mathbb{E}[W_N(D(1/p), RS(D(\rho))]$, which is computed by the recursion. We choose N suitably large so that $\mathbb{E}[W_N]$ is close to $\mathbb{E}[W]$, which can be seen when the values of several successive N change only negligibly.

7.1 Performance of the negative binomial algorithm

We set different truncation levels K and N to study the computational accuracy and effort of Algorithm 1. In the experiment, we consider the case $c_a^2 = 4.0$, so that $p = 1/(1 + c_a^2) = 0.2$. We fix the truncation level $N = 10^3$ and let K vary from 10^3 to 8×10^3 to execute Algorithm 1. (It is good to have k in the outer loop because $p = 1/(1+c_a^2) < 0.5$.) The results are shown in Table 2 for a range of traffic intensities from $\rho = 0.10$ to $\rho = 0.99$. Also shown for comparison in the last two columns are the simulation estimates from the highly accurate Minh-Sorli [17] simulation method, as given in Table 6.

For $\rho \leq 0.90$, the recursive algorithm with truncation level $N = 10^3$ and $K = 3 \times 10^3$ performs well, but for $\rho \geq 0.95$, the numerical values of $\mathbb{E}[W]$ converge as K increases but are not close to the simulation results.

Table 2 Performance of Algorithm 1 for $p = 1/(1 + c_a^2) = 0.2$ with different truncation levels

$\rho \backslash K$	Algorithm Procedure 1 with $N = 10^3$				Minh and Sorli Algorithm	
	1×10^3	2×10^3	4×10^3	8×10^3	$T = 1 \times 10^7$	95%CI
0.1	0.422229	0.422229	0.422229	0.422229	0.422	7.79E-05
0.2	0.903885	0.903885	0.903885	0.903885	0.904	1.30E-04
0.3	1.499234	1.499234	1.499234	1.499234	1.499	1.71E-04
0.4	2.304105	2.304105	2.304105	2.304105	2.304	1.90E-04
0.5	3.470132	3.470132	3.470132	3.470132	3.470	2.25E-04
0.6	5.294825	5.294825	5.294825	5.294825	5.294	2.43E-04
0.7	8.441305	8.441305	8.441305	8.441305	8.442	3.05E-04
0.8	14.916481	14.916937	14.916937	14.916937	14.917	3.22E-04
0.9	34.276662	34.673925	34.718140	34.718140	34.722	5.17E-04
0.95	66.874413	71.232241	73.264743	73.264743	74.621	7.11E-04
0.98	139.659440	152.638886	162.915010	162.915010	194.556	9.29E-04
0.99	245.012809	262.661919	278.499123	278.499123	394.532	1.45E-03

7.2 Refinement to the negative binomial algorithm for heavy traffic

The difficulty in heavy traffic occurs because as ρ increases, we need larger values of n. For extremely large n, as is needed in heavy traffic, p^n and $(1-p)^n$ are eventually very small numbers. That causes the probability to become too small to be represented in the implemented floating-point number system. Hence, in heavy traffic, the basic recursive algorithm broke down because the large values of n caused underflow.

As when computing the steady-state of the birth-and-death processes, for example, as in Sect. 7 of [28], for very large n we can encounter underflow problems if we start the recursion at 0, but it can be avoided by starting the recursion elsewhere. We avoid the underflow problem by doing two recursions, one up and the other down, starting from the mean.

It now remains to consider how to do the truncations. First, consider the truncation of the sum on k for given n. For given n,

$$\mathbb{E}[NB(n,1-p)] \equiv m(n) = \frac{n(1-p)}{p} \quad \text{and}$$
$$Var(NB(n,1-p)) \equiv \sigma^2(n) = \frac{n(1-p)}{p^2}. \tag{7.6}$$

From the central limit theorem, we know that the NB distribution is approximately Gaussian with a mean near its mode. In particular,

$$NB(n,1-p) \approx \mathcal{N}(m(n),\sigma^2(n)) \quad \text{as} \quad n \to \infty. \tag{7.7}$$

for $m(n)$ and $\sigma^2(n)$ in (7.6). Hence, for large n, it suffices to consider only a modest range of k, i.e., of order $O(\sqrt{n})$. In particular, it should suffice to consider $m(n) - a\sigma(n) \le k \le m(n) + a\sigma(n)$ for, for example, $a = 10, 20$. However, we need to add a term for small k. For $k \le m(n) - a\sigma(n)$, we let $\mathbb{P}(NB(n,1-p) > k) = 1$. That means we add $(m(n) - a\sigma(n)) \vee 0$, where $a \vee b \equiv \max\{a,b\}$.

Finally, the relevant values of n depend on the traffic intensity ρ and other model parameters. For heavy traffic (large ρ), we can use the approximation (2.9) to estimate the relevant n. Moreover, given that the heavy-traffic limit of the waiting-time distribution is exponential, we can see the relevant range of n.

Suppose N is the upper bound of n. As a consequence, for large N, we consider $k \in [\max(m(n) - 20\sqrt{N}, 0), m(n) + 20\sqrt{N}]$ in the implementation. Here is how we proceed: For fixed $n \le N$, we start from the mean in (7.5) and let $\mathbb{P}(NB(n,1-p) = n(1-p)/p) = 1$ and then do recursive formula (7.5) up and down separately. Define the mean $n(1-p)/p$ by $m(n)$. The two-part recursion going up and down becomes

$$\mathbb{P}(NB(n,1-p) = m(n)+j) = \frac{\mathbb{P}(NB(n,1-p) = m(n)+j-1)(n+m(n)+j-1)(1-p)}{m(n)+j}, \tag{7.8}$$

$$\mathbb{P}(NB(n,1-p) = m(n)-j) = \frac{\mathbb{P}(NB(n,1-p) = m(n)-j+1)(m(n)-j+1)}{(n+m(n)-j)(1-p)} \tag{7.9}$$

for $j \geq 1$. Afterward, we normalize the values obtained from the above recursion to get probabilities $\mathbb{P}(NB(n, 1-p) = k)$ for any k given n.

As in Algorithm 1, in Algorithm 2 we apply Theorem 3.1 to obtain an efficient algorithm for computing the UB $\mathbb{E}[W(F_0, G_{u^*})]$. Thus, we initialize by the constant term that depends only on the parameter vector $(1, c_a^2, \rho, c_s^2)$.

Algorithm 2 negative binomial recursion (up and down from the mean)

Initially set $\mathbb{E}[W] \leftarrow \rho c_a^2 + \frac{\rho^2 c_s^2}{2(1-\rho)}$, $p = (1 + c_a^2)^{-1}$, and $m(n) = n(1-p)/p$.
for $n \in [1, N]$ **do**
  $nbpdf(1, m(n)) \leftarrow 1$
  for $k \in [m(n) - 20\sqrt{N} + 1, m(n)]$ **do**
    $nbpdf(1, k-1) \leftarrow nbpdf(1, k)/(n+k-1)(k)/(1-p)$
  for $k \in [m(n), m(n) + 20\sqrt{N} - 1]$ **do**
    $nbpdf(1, k+1) \leftarrow nbpdf(1, k)(n+k)/(k+1)(1-p)$
  Normalize $nbpdf$ to obtain $\mathbb{P}(NB(n, 1-p) = k)$
  $S(n) \leftarrow \sum_k \mathbb{P}(NB(n, 1-p) = k) \max((n+k)\rho - n/p, 0)$
  $\mathbb{E}[W] \leftarrow \mathbb{E}[W] + S(n)/n$
Output $\mathbb{E}[W]$

We now carefully compare the negative binomial pmf values generated from the basic recursion (7.5) used in Algorithm 1 with the values obtained in the new up–down recursion used in Algorithm 2 in Table 3. We focus on the terms after $m(n)$ and report the values from the term $m(n)$ to $m(n) + 10$.

For $n \leq 10^2$, the results from the two methods agree to all digits shown, but a significant difference occurs when $n = 10^3$. At $n = 10^3$, underflow occurs in Algorithm 1, which causes the errors we saw for large ρ in Table 2.

7.3 Performance studies for the refined negative binomial algorithm

Table 4 shows that Algorithm 2 is also very efficient for $\rho \leq 0.95$. Table 4 shows that the new algorithm is effective if we increase N from 10^3 to 10^4 as ρ increases.

In particular, the numerical algorithm is more efficient than the simulation. It requires no more than 30 seconds CPU time in the worse case ($N = 2 \times 10^4$, $\rho = 0.95$) to produce more than ten decimal places accuracy, while the MS simulation algorithm only attains 10^{-4} confidence interval level for $0.5 \leq \rho \leq 0.95$ while producing three decimal places accuracy within around 30 seconds CPU time.

Next, we apply Algorithm 2 for the heavy-traffic cases with $\rho = 0.98$ and $\rho = 0.99$. To do so, we restrict the range of k to $k \leq m(n) + 20\sqrt{N}$ for the purpose of setting smaller N. Table 5 shows that the poor performance of the NB algorithm in Table 2 has been improved dramatically by the alternative algorithm.

Remark 7.1 *(suggested parameters)* Our experiments suggest that, for typical values of p (not too small), it suffices to set $N = O(1/(1-\rho)^3)$ to obtain highly accurate results.

Table 3 Comparison of the basic and up–down recursions for generating values of the negative binomial pmf in Algorithms 1 and 2 for the case $p = 0.2$

k	Alg 1 $n = 10$	Alg 2 $n = 10$	k	Alg 1 $n = 10^2$	Alg 2 $n = 10^2$	k	Alg 1 $n = 10^3$	Alg 2 $n = 10^3$
40	0.0279638	0.0279638	400	0.0089128	0.0089128	4000	0	0.0028207
42	0.0265023	0.0265023	402	0.0088641	0.0088641	4002	0	0.0028192
44	0.0247071	0.0247071	404	0.0087983	0.0087983	4004	0	0.0028170
46	0.0226875	0.0226875	406	0.0087160	0.0087160	4006	0	0.0028144
48	0.0205443	0.0205443	408	0.0086179	0.0086179	4008	0	0.0028111
50	0.0183647	0.0183647	410	0.0085047	0.0085047	4010	0	0.0028074

Table 4 Performance of Algorithm 2 for $p = 0.2$ with different truncation levels

$\rho \backslash N$	Algorithm 2					Minh and Sorli Algorithm	
	2×10^3	4×10^3	8×10^3	1.6×10^4	2×10^4	$T = 1 \times 10^7$	95%CI
0.1	0.422229	0.422229	0.422229	0.422229	0.422229	0.422	7.79E-05
0.2	0.903885	0.903885	0.903885	0.903885	0.903885	0.904	1.30E-04
0.3	1.499234	1.499234	1.499234	1.499234	1.499234	1.499	1.71E-04
0.4	2.304105	2.304105	2.304105	2.304105	2.304105	2.304	1.90E-04
0.5	3.470132	3.470132	3.470132	3.470132	3.470132	3.470	2.25E-04
0.6	5.294825	5.294825	5.294825	5.294825	5.294825	5.294	2.43E-04
0.7	8.441305	8.441305	8.441305	8.441305	8.441305	8.442	3.05E-04
0.8	14.916937	14.916937	14.916937	14.916937	14.916937	14.917	3.22E-04
0.9	34.721476	34.721484	34.721484	34.721484	34.721484	34.722	5.17E-04
0.95	74.552341	74.619631	74.620917	74.620937	74.620937	74.621	7.11E-04

Table 5 Performance of Algorithm 2 in heavy traffic, for $p = 0.2$

Case	Algorithm 2				Minh and Sorli	
$\rho \backslash N$	1×10^4	2×10^4	3×10^4	4×10^4	$T = 1 \times 10^7$	
0.98	194.0544167173	194.5385548017	194.5559125683	194.5567071265	194.556	9.29E-04
	5×10^4	1×10^5	2×10^5	3×10^5		
0.98	194.5567179973	194.5567742874	194.5567742874	194.5567742874	194.556	9.29E-04
Case	Algorithm 2				Minh and Sorli	
$\rho \backslash N$	1×10^4	3×10^4	5×10^4	1×10^5	$T = 1 \times 10^7$	
0.99	372.0880005430	372.0880005430	391.8858614678	394.5238008176	394.532	1.45E-03
	2×10^5	3×10^5	4×10^5	5×10^5		
0.99	394.5331823499	394.5331886695	394.5331886695	394.5331886695	394.532	1.45E-03

Remark 7.2 *(opportunity for simulation efficiency)* Since the service-time variability parameter c_s^2 is not used to evaluate $\mathbb{E}[W(D(1/p), RS(D(\rho), p)]$ in Algorithm 2, Tables 4 and 5 can be reused to compute $\mathbb{E}[W(F_0, G_{u^*})]$ with any other c_s^2 via Theorem 3.1.

8 Exploiting the idle-time representation

To develop alternative algorithms, following [16,17] and [29], we relate the mean waiting time given the first two moments of the interarrival time and service time to the first two moments of the idle time I. In Sect. 8.1, we review the basic relation. In Sect. 8.2, we discuss the implications of the relation when we let $M_s \to \infty$. In Sect. 8.3, we show the advantage of combining Theorem 8.1 and Corollary 5.1. Later, in Sect. 9.2, we apply the representation to develop a new numerical algorithm based on computing absorption probabilities in finite-state discrete-time Markov chains (DTMCs).

8.1 The basic representation

The key relation is in the following theorem.

Theorem 8.1 (the idle-time representation, Theorem 1 of Marshall [16]) *In the $GI/GI/1$ queue with cdf's F and G having parameter 4-tuple $(1, c_a^2, \rho, c_s^2)$,*

$$\mathbb{E}[W] = \psi(1, c_a^2, \rho, c_s^2) - \phi(I), \tag{8.1}$$

where

$$\psi(1, c_a^2, \rho, c_s^2) \equiv \frac{\mathbb{E}[(U-V)^2]}{2\mathbb{E}[U-V]} = \frac{\rho^2([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)} + \frac{1-\rho}{2} \tag{8.2}$$

and

$$\phi(I) = \frac{\mathbb{E}[I^2]}{2\mathbb{E}[I]} = \mathbb{E}[I_e], \tag{8.3}$$

with I being the steady-state idle time and I_e being a random variable with the associated stationary excess distribution (as in renewal theory).

Notice that $\mathbb{E}[W]$ depends on the model distributions F and G beyond the parameter vector $(1, c_a^2, \rho, c_s^2)$ only through $\phi(I) = \mathbb{E}[I_e]$ in (8.3). For the $M/GI/1$ model, I is distributed as F, $\phi(I) = 1$ and simple algebra yields the exact Pollaczek-Khintchine formula. In general, the first term on the right in (8.2) is the [14] upper bound. For the [14] bound to be obtained, the second term on the right in (8.2) would have to be exactly canceled by the second term on the right in (8.1).

8.2 The limit as $M_s \to \infty$

This section is based on the notion that the upper bound is obtained as the limit of $\mathbb{E}[W]$ within the $F_0/G_u/1$ model as $M_s \to \infty$. Because the mean waiting time is not continuous as $M_s \to \infty$, but the idle-time distribution is, we approach the upper bound via the idle time.

We can apply Theorems 4.1 and 8.1 to obtain a limit within the decomposition. For that purpose, let $\phi(I; U, V)$ denote $\phi(I)$ in (8.3) for the model with interarrival time U and service time V. We will consider $U = D(1/p)$ and $V = RS(D(\rho), p)$.

Theorem 8.2 (limit within the decomposition) *For the $F_0/G_u/1$ model with parameter vector $(1, c_a^2, \rho, c_s^2)$ and service-distribution support $[0, \rho M_s]$,*

$$\lim_{M_s \to \infty} \mathbb{E}[W(F_0, G_u)] = \psi(1, c_a^2, \rho, c_s^2) + \rho c_a^2 - \phi(I; (1 + c_a^2), 0, \rho(1 + c_a^2), \bar{c}_s^2), \tag{8.4}$$

where $\phi(I; D(1/p), RS(D(\rho), p))$ means (8.3) for the $D(1/p)/RS(D(\rho), p))/1$ model and the parameter vector for that model is $((1 + c_a^2), 0, \rho(1 + c_a^2), \bar{c}_s^2)$ for $\bar{c}_s^2 \equiv c_a^2/(1 + c_a^2)$.

Proof We apply Theorems 4.1 and 8.1 to write

$$\lim_{M_s \to \infty} \mathbb{E}[W(F_0, G_u)] = \psi(1, c_a^2, \rho, c_s^2) - \phi(I; 1, c_a^2, \rho, 0), \tag{8.5}$$

where

$$\psi(1, c_a^2, \rho, c_s^2) \equiv \frac{c_a^2 + \rho^2 c_s^2}{2(1 - \rho)} + \frac{1 - \rho}{2}, \tag{8.6}$$

which is independent of M_s and thus is unchanged by the limit on M_s. However, the second term changes, consistent with the distribution G_u approaching $D(\rho)$ as $M_s \to \infty$, and having the limiting mean but 0 variance. As a consequence,

$$\begin{aligned} \lim_{M_s \to \infty} \mathbb{E}[W(F_0, G_u)] &= \psi(1, c_a^2, \rho, c_s^2) + \rho c_a^2 - \lim_{M_s \to \infty} \phi(I) \\ &= \psi(1, c_a^2, \rho, c_s^2) + \rho c_a^2 - \phi(I; D(1/p), RS(D(\rho), p)), \\ &= \psi(1, c_a^2, \rho, c_s^2) + \rho c_a^2 \\ &\quad - \phi(I; (1 + c_a^2), 0, \rho(1 + c_a^2), \bar{c}_s^2). \end{aligned} \tag{8.7}$$

□

Theorem 8.2 implies that it only remains to evaluate the idle-time term $\phi(I)$ in (43) as it arises in the last line of (8.7) for the $D(1/p)/RS(D(\rho), p)/1$ model, for which the only randomness is in the random sum in the service times. The random sum is a geometric random sum of constants in this case. When we apply the Minh-Sorli

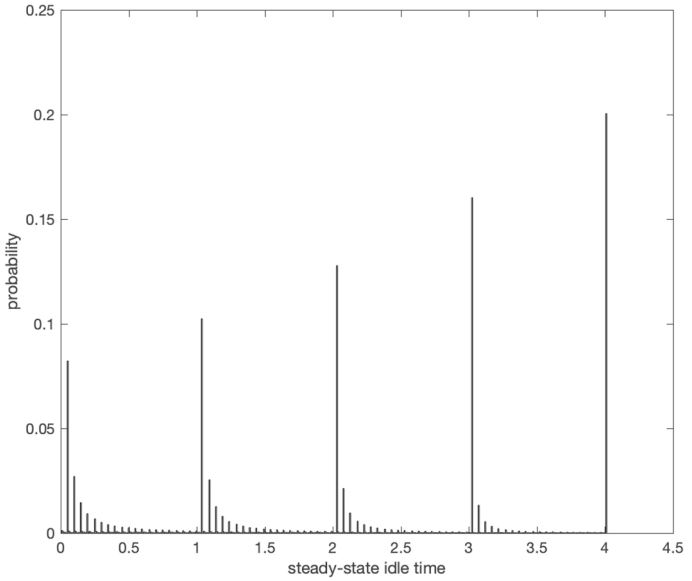

Fig. 1 Simulation estimates of the steady-state idle-time distribution in the $F_0/D/1$ model and the $F_0/G_u/1$ model as $M_s \to \infty$ under traffic level $\rho = 0.99$ and $c_a^2 = c_s^2 = 4.0$

[17] method for simulation, it suffices to reduce variance by ignoring the large M_s. We treat the service times as D with mean ρ. But, when we do so, we have to make adjustments in the final formulas as indicated above.

To illustrate the algorithm for computing $\mathbb{E}[W(F_0, G_{u^*})]$ in (2.10) by using the idle-time representation, and because it is directly interesting, we present a simulation estimate of the idle-time distribution for $\rho = 0.99$ and $c_a^2 = c_s^2 = 4.0$ in Fig. 1. We remark that this is also the steady-state idle-time distribution for model $F_0/D/1$.

8.3 Combining Theorem 8.1 and Corollary 5.1

Combining Theorem 8.1 and Corollary 5.1, we obtain

Corollary 8.1 (reduction to idle time) *For the $GI/GI/1$ model with extremal interarrival-time cdf F_0, extremal service-time cdf G_u and G_{u*} defined in (2.10),*

$$\mathbb{E}[W(F_0, G_{u^*})] = \frac{c_a^2 + \rho^2 c_s^2}{2(1-\rho)} + \frac{1-\rho}{2} - \phi(I; 1, c_a^2, \rho, c_s^2), \tag{8.8}$$

where I is the idle time in an $F_0/G_{u^*}/1$ queue or, equivalently, in a $F_0/D/1$ queue for an appropriate D.

Corollary 8.1 shows that to determine the UB $\mathbb{E}[W(F_0, G_{u^*})]$ defined in (2.10), it suffices to calculate the term $\phi(I; 1, c_a^2, \rho, c_s^2)$ in (8.3) for the $F_0/D/1$ model via effective algorithms. In contrast, Theorem 8.2 concludes that it suffices to calculate ϕ in (8.3) for the $D(1/p)/RS(D(\rho), p)/1$ model, but we see that these are equivalent,

because we can go from one to the other by applying Theorem 4.1. Thus, we conclude that Sects. 4 and 5 are two different ways to reach essentially the same conclusion.

9 Simulation algorithms and experiments

In this section, we compare three different simulation algorithms for estimating the extremal mean steady-state waiting time $\mathbb{E}[W(F_0, G_{u^*})]$: (i) the standard Monte Carlo (MC) algorithm, (ii) the Minh-Sorli [17] (MS) algorithm and (iii) the method from Sect. 9.2.2 based on simulating a discrete-time random walk (RW). We now describe the [17] simulation algorithm.

9.1 The Minh-Sorli [17] simulation algorithm

The idea is to exploit Theorem 8.1. In particular, we exploit the discrete event simulation method to estimate the first two moments of the steady-state idle period I; i.e., we exploit (8.1) and estimate $\phi(I)$ in (8.3). In the simulation algorithm, the successive events are classified in three ways: (i) arrival is next, (ii) departure is next and (iii) next event occurs after given time T, where T is total simulation length.

Thus, within each replication, we estimate $\mathbb{E}[I]$ and $\mathbb{E}[I^2]$ and then apply Theorem 8.1 to obtain an associated estimate of $\mathbb{E}[W]$. We then compute confidence intervals for this alternative estimate of $\mathbb{E}[W]$ by performing multiple replications, as described in the online supplement.

9.2 An idle-time random walk simulation algorithm

Theorem 8.2 implies that $\mathbb{E}[W(F_0, G_{u^*})]$ in (2.10) can be expressed in terms of the first two moments of the steady-state idle time I in the $D(1/p)/RS(D(\rho), p)/1$ model and the parameter vector $(1, c_a^2, \rho, c_s^2)$. In this section, we show how to develop algorithms to calculate the distribution and moments of I in the $D(1/p)/RS(D(\rho), p)/1$ model based on a random walk absorption representation.

9.2.1 A random walk absorption representation of the idle time

For the reduced model $D(1/p)/RS(D(\rho), p)/1$, the steady-state idle time can be expressed in terms of a random walk $\{Y_k : k \geq 0\}$ defined in terms of the recursion

$$Y_{k+1} = Y_k + \rho N_k - (1 + c_a^2), \quad k \geq 1, \quad Y_0 \equiv 0, \tag{9.1}$$

where N_k is a negative binomial random variable with parameter p on the integers, while $1 + c_a^2 = 1/p$ is a deterministic interarrival time. Hence, $\{N_k : k \geq 1\}$ is an i.i.d. sequence with $N_k \stackrel{\mathrm{d}}{=} RS(D(1), p)$. The random variables $\rho N_k - (1 + c_a^2)$ in (9.1) are the steps of the random walk. Each step is the net input of work from one arrival time to the next. Because N_k takes values on the positive integers, the possible steps are $k\rho - (1 + c_a^2)$ for $k \geq 1$, so that $\rho N_k - (1 + c_a^2) \geq \rho - (1 + c_a^2)$.

As long as $Y_k \geq 0$, Y_k represents the work in the system at the time of the k^{th} arrival, starting empty. The number of customers served in that busy cycle, N_c, and the length of a busy cycle, C, are then

$$N_c = \inf\{k \geq 1 : Y_k \leq 0\} \quad \text{and} \quad C = N_c(1 + c_a^2). \tag{9.2}$$

The associated idle-time random variable is distributed as

$$I \stackrel{\text{d}}{=} -Y_{N_c}, \quad \text{so that} \quad 0 \leq I \leq c_a^2 + 1 - \rho. \tag{9.3}$$

9.2.2 An idle-time simulation algorithm

Given N i.i.d. copies of I, each obtained via (9.1)-(9.3), we can estimate the cdf $F_I(x) \equiv \mathbb{P}(I \leq x)$, $x \geq 0$, by the empirical cdf

$$\bar{F}_I(x) \equiv N^{-1} \sum_{i=1}^{N} I(I_i \leq x). \tag{9.4}$$

To estimate the p^{th} moment $\mathbb{E}[I^p]$, we can compute the sample mean, using

$$\bar{I}_N \equiv R^{-1} \sum_{i=1}^{R} N^{-1} \sum_{i=1}^{N} I_i^p, \tag{9.5}$$

where R is the number of replications.

9.3 Comparison of the three simulation algorithms

We now apply and compare our three simulation algorithms to estimate $\mathbb{E}[W(F_0, G_{u^*})]$ in (2.10): (i) the standard Monte Carlo (MC) algorithm, (ii) the Minh-Sorli [17] (MS) algorithm and (iii) the method from Sect. 9.2.2 based on simulating a discrete-time random walk.

Estimates of $\mathbb{E}[W(F_0, G_{u^*})]$ by the three algorithms are shown in Table 6. These are for the case $c_a^2 = c_s^2 = 4.0$ and $M_s = 1000$ for the MC algorithm and $M_s = \infty$ for the other two simulation algorithms. Results are reported for a range of traffic intensities ranging from $\rho = 0.1$ to $\rho = 0.99$.

We now describe the simulation parameters for each algorithm. The MC method has truncation level $N = 10^7$ in sample mean and we make $R = 20$ i.i.d replications. The MS method has total run length $T = 10^6$ again with $R = 20$ i.i.d replications. (We use all idle periods that fall within that time interval.)

Table 6 shows the simulation estimates from all three approaches. Table 6 shows that the simulation methods are mutually confirming, but that the confidence intervals are quite different. The accuracy is ordered by $MS > RW > MC$ with MS being best. For additional details, see the online supplement.

Table 6 Comparison of Three Different Simulation Algorithms

Simulation estimates of $\mathbb{E}[W(F_0, G_{u^*})]$ for $c_a^2 = c_s^2 = 4$						
ρ	MC UB	95% CI Length	MS UB	95% CI Length	RW UB	95% CI Length
0.10	0.422	5.08E-04	0.422	7.79E-05	0.422	9.28E-04
0.20	0.904	2.29E-03	0.904	1.30E-04	0.903	1.64E-03
0.30	1.484	4.44E-03	1.499	1.71E-04	1.498	1.47E-03
0.40	2.310	1.47E-02	2.304	1.90E-04	2.305	1.68E-03
0.50	3.472	2.15E-02	3.470	2.25E-04	3.472	2.00E-03
0.60	5.276	5.39E-02	5.294	2.43E-04	5.295	3.14E-03
0.70	8.381	7.80E-02	8.442	3.05E-04	8.442	2.62E-03
0.80	15.016	1.54E-01	14.917	3.22E-04	14.919	3.13E-03
0.90	34.525	4.60E-01	34.722	5.17E-04	34.720	1.95E-03
0.95	76.059	1.24E+00	74.621	7.11E-04	74.621	2.26E-03
0.98	193.206	3.07E+00	194.556	9.29E-04	194.558	2.75E-03
0.99	394.763	1.02E+01	394.532	1.45E-03	394.532	2.62E-03

10 Conclusions

In this paper, we developed numerical and simulation algorithms to compute the widely conjectured tight upper bound for the mean steady-state waiting time $\mathbb{E}[W]$ in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions, as specified by the parameter vector $(1, c_a^2, \rho, c_s^2)$. It is conjectured that this tight bound is attained asymptotically be two-point distributions, specifically by the pair (F_0, G_u) defined in Sect. 2.3.2 as $M_s \to \infty$.

Our algorithms are based on an explicit representation for $\mathbb{E}[W(F_0, G_{u^*})]$ in terms of $\mathbb{E}[W(D(1/p), RS(D(\rho), p))]$ in Theorem 3.1, where G_{u^*} is defined in (2.10) and RS denotes a geometric random sum. Theorem 3.2 gives a convenient explicit formula for an UB to $\mathbb{E}[W(F_0, G_{u^*})]$. Table 1 shows that the UB formula is very accurate and that the new results provide significant improvement over previous bounds.

In Sect. 6, we derived an explicit expression for $\mathbb{E}[W(D(1/p), RS(D(\rho), p))]$ in some special cases, which yields an explicit expression for $\mathbb{E}[W(F_0, G_{u^*})]$ in those cases. In Sect. 7, we developed effective numerical algorithms to compute the mean steady-state waiting time $\mathbb{E}[W(D(1/p), RS(D(\rho), p))]$ using recursive algorithms for the negative binomial probability mass function. We also conducted experiments showing that the algorithms are effective. We exposed and resolved an underflow problem that can arise in heavy traffic.

In Sect. 8, using the Minh-Sorli [17] insight, we showed that it also suffices to compute the first two moments of the steady-state idle-time distribution in the $D(1/p)/RS(D(\rho), p)/1$ model. Theorem 8.2 shows that the idle time is better behaved than the waiting time as the extremal service mass increases. In Sect. 9 and the online supplement [5], we studied three possible simulation algorithms for estimating $\mathbb{E}[W(F_0/G_{u^*})]$: the standard Monte Carlo simulation (MC) and two methods exploiting the idle-time representation: the Minh-Sorli [17] (MS) algorithm and a new

algorithm based on a discrete-time random walk (RW). We showed that both MS and RW provide significant improvement over MC, but that MS tends to be best.

Overall, we found that, first, the reductions are powerful for simplifying the algorithms and, second, that the refined negative binomial numerical algorithm in Sect. 7 and the [17] simulation algorithm in Sect. 9 are most effective for computing $\mathbb{E}[W(D(1/p), RS(D(\rho), p))]$.

Finally, there are many important directions for further research, including providing a proof that $\mathbb{E}[W(F_0, G_{u^*})]$ in (2.10) and Theorem 3.1 does indeed provide an upper bound. It also remains to consider additional properties of the cdf's F and G that will narrow the range of possible values, as was done in [12,15,25] and [26] for the $GI/M/1$ model. Some new contributions appear in [6].

Acknowledgements This research was supported by NSF CMMI 1634133.

References

1. Abate, J., Whitt, W.: The Fourier-series method for inverting transforms of probability distributions. Queueing Syst. **10**, 5–88 (1992)
2. Abate, J., Choudhury, G.L., Whitt, W.: Calculation of the GI/G/1 steady-state waiting-time distribution and its cumulants from Pollaczek's formula. Archiv fur Elektronik und Ubertragungstechnik **47**(5/6), 311–321 (1993)
3. Asmussen, S.: Applied Probability and Queues, second edn. Springer, New York (2003)
4. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1999)
5. Chen, Y., Whitt, W.: Supplement to Algorithms for the Upper Bound Mean Waiting Time in the $GI/GI/1$ Queue. Columbia: Columbia University, http://www.columbia.edu/~ww2040/allpapers.html, (2019)
6. Chen, Y., Whitt, W.: Set-Valued Queueing Approximations Given Partial Information. Columbia: Columbia University, http://www.columbia.edu/~ww2040/allpapers.html (2019)
7. Chen, Y., Whitt, W.: Extremal $GI/GI/1$ Queues Given Two Moments: Exploiting Tchebycheff Systems. Columbia: Columbia University, http://www.columbia.edu/~ww2040/allpapers.html, (2020)
8. Chung, K.L.: A Course in Probability Theory, third edn. Academic Press, New York (2001)
9. Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., Knuth, D.E.: On the Lambert w function. Adv. Comput. Math. **5**, 329–359 (1996)
10. Daley, D.J.: Inequalities for moments of tails of random variables, with queueing applications. Zeitschrift fur Wahrscheinlichkeitsetheorie Verw. Gebiete **41**, 139–143 (1977)
11. Daley, D.J., Kreinin, A Ya., Trengove, C.D.: Inequalities concerning the waiting-time in single-server queues: a survey. In: Bhat, U.N., Basawa, I.V. (eds.) Queueing and Related Models, pp. 177–223. Clarendon Press, Oxford (1992)
12. Eckberg, A.E.: Sharp bounds on Laplace–Stieltjes transforms, with applications to various queueing problems. Math. Oper. Res. **2**(2), 135–142 (1977)
13. Halfin, S.: Batch delays versus customer delays. Bell Lab. Tech. J. **62**(7), 2011–2015 (1983)
14. Kingman, J.F.C.: Inequalities for the queue $GI/G/1$. Biometrika **49**(3/4), 315–324 (1962)
15. Klincewicz, J.G., Whitt, W.: On approximations for queues, II: shape constraints. AT&T Bell Lab. Tech. J. **63**(1), 139–161 (1984)
16. Marshall, K.T.: Some inequalities in queueing. Oper. Res. **16**(3), 651–668 (1968)
17. Minh, D.L., Sorli, R.M.: Simulating the $GI/G/1$ queue in heavy traffic. Oper. Res. **31**(5), 966–971 (1983)
18. Muller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley, New York (2002)
19. Ott, T.J.: Simple inequalities for the $D/G/1$ queue. Oper. Res. **35**(4), 589–597 (1987)
20. Ross, S.M.: Stochastic Processes, second edn. Wiley, New York (1996)
21. Ross, S.M.: Introduction to Probability Models, eleventh edn. Academic Press, New York (2014)
22. Stoyan, D.: Comparison Methods for Queues and Other Stochastic Models. Wiley: New York, 1983. Translated and edited from 1977 German Edition by D. J. Daley (1977)

23. Stoyan, D., Stoyan, H.: Inequalities for the mean waiting time in single-line queueing systems. Eng. Cybern. **12**(6), 79–81 (1974)
24. Whitt, W.: Comparing batch delays and customer delays. Bell Lab. Tech. J. **62**(7), 2001–2009 (1983)
25. Whitt, W.: On approximations for queues, I: extremal distributions. AT&T Bell Lab. Tech. J. **63**(1), 115–137 (1984)
26. Whitt, W.: On approximations for queues, III: mixtures of exponential distributions. AT&T Bell Lab. Tech. J. **63**(1), 163–175 (1984)
27. Whitt, W.: Deciding which queue to join: some counterexamples. Oper. Res. **34**(1), 55–62 (1986)
28. Whitt, W.: Engineering solution of a basic call-center model. Manag. Sci. **51**, 221–235 (2005)
29. Wolff, R.W., Wang, C.: Idle period approximations and bounds for the $GI/G/1$ queue. Adv. Appl. Probab. **35**(3), 773–792 (2003)